MA 421 Prerequisite Review - Calculus

This document is written to be a review of the calculus relevant to linear programming and optimization rather than a full calculus lecture note. Please reach out if you spot any typos in this document.

---

1 Single-Variable Differentiation

We start this with a review of differentiation with respect to one variable. Per the definition, derivatives measure the rate of change of a function and are very important to identity maxima and minima of functions.

Derivatives

Let
$$f : \mathbb{R} \to \mathbb{R}$$
The derivative of f at a point x is defined by
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Geometrically, $f'(x)$ represents the slope of the tangent line to the function f at the point x. The sign of the derivative tells us about the local behavior of the function.

- If $f'(x) > 0$, then f is locally increasing
- If $f'(x) < 0$ then f is locally decreasing
- If $f'(x) = 0$, then f may be a local minimum, local maximum, or neither.

For this course, you should be comfortable finding derivatives using the product rule, quotient rule, and chain rule.

Critical Points

A point x^* is a *critical point* of f if
$$f'(x^*) = 0$$
or if $f'(x^*)$ does not exist

Critical points are very important in optimization because local minima and maxima often occur at critical points.

Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable at x_0, and suppose that f has a local minimum or local maximum at x_0, then $f'(x_0) = 0$.

This is called the *first-order necessary condition* for optimality. To prove this, start by assuming that f has a local minimum or maximum at x_0 and then use the limit definition of the derivative to complete the proof.

One thing that is important to remember is that this condition is necessary, but not sufficient. This means that

$$f'(x^*) = 0$$

does not directly tell us that x^* is a maximum or minimum.

Example. Let us consider $f(x) = x^3$

This gives us $f'(x) = 3x^2$ so $f'(0) = 0$. But $x = 0$ is neither a local maximum nor a local minimum for x^3

The second derivative

The second derivative

$$f''(x)$$

describes how the first derivative is changing and tells us about the *curvature* of the function.

Suppose that x^* is a critical point satisfying $f'(x^*) = 0$. Then the second derivative test gives us that

- If
  $$f''(x^*) > 0$$
  then x^* is a local minimum
- If
  $$f''(x^*) < 0$$
  then x^* is a local maximum
- If
  $$f''(x^*) = 0$$
  then the test is inconclusive

Example

Let us consider

$$f(x) = x^3 - 3x^2 + 2$$

First, we compute the derivative which gives us

$$f'(x) = 3x^2 - 6x = 3x(x - 2)$$

To find the critical points, we set $f'(x) = 0$ which gives us

$$f'(x) = 3x(x - 2) = 0$$

So the critical points are $x = 0$ and $x = 2$

Then we compute the second derivative. This is

$$f''(x) = 6x - 6$$

At $x = 0$, the second derivative is $f''(0) = -6 < 0$, therefore $x = 0$ is a local maximum.

At $x = 2$, the second derivative is $f''(2) = 6 > 0$, therefore $x = 2$ is a local minimum.

The graph of the function below confirms the results we derived analytically.

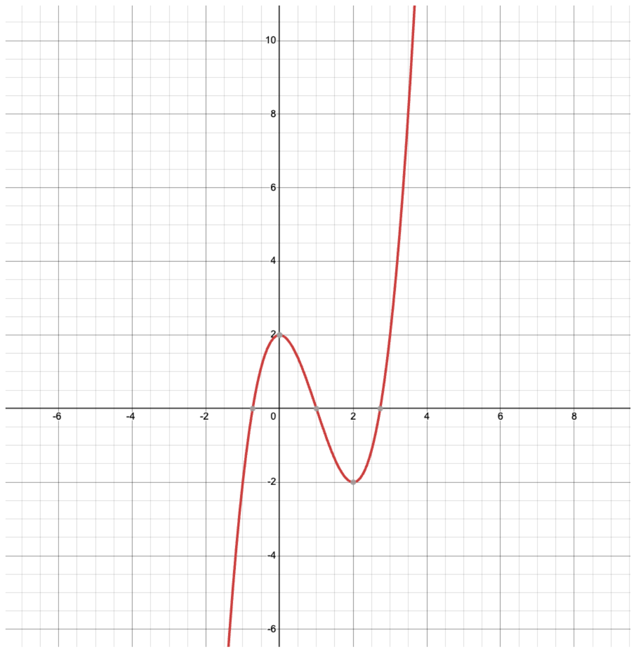

Figure 1: Graph of $f(x) = x^3 - 3x^2 + 2$

2 Multivariable Differentiation

In this course, we study functions of several variable so it is important to extend the ideas from single variables. When we study nonlinear optimization, we will often work with functions of the form

$$f : \mathbb{R}^n \to \mathbb{R}$$

In this case, instead of working with a single derivative, we will use partial derivatives, gradients, and Hessian matrices to understand how these functions behave.

Partial Derivatives

Suppose that

$$f : \mathbb{R}^n \to \mathbb{R}$$

and let

$$x = (x_1, x_2, \ldots, x_n)$$

The partial derivative of f with respect to each x_i measures how f changes as we keep changing x_i while keeping the other variables fixed. It is defined by

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}$$

provided this limit exists.

Example. Let us consider $f(x, y) = x^2 + 3xy + y^2$

Then we have

$$\frac{\partial f}{\partial x} = 2x + 3y$$

and

$$\frac{\partial f}{\partial y} = 3x + 2y$$

When computing $\frac{\partial f}{\partial x}$, we treat y as a constant and when computing $\frac{\partial f}{\partial y}$, we treat x as a constant.

The Gradient

The partial derivatives of a function are collected into a vector called the *gradient.*

Let $f : \mathbb{R}^n \to \mathbb{R}$. The gradient of f at a point x is

$$\nabla f(x) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix}$$

The gradient is the multi-variable equivalent of the derivative from single variable calculus. It contains information about how the function changes with respect to each of its variables.

One of the most classic optimization algorithms, used a lot in Machine Learning especially is called **Gradient Descent**. Read more about it here If you are interested.

The gradient also has a very important geometric interpretation. At a point x, the gradient $\nabla f(x)$ points in the direction of steepest increase of the function. As a result $-\nabla f(x)$ points in the direction of steepest decrease.

Directional Derivatives

In one variable, the derivative tells us how a function changes as we move along the plane. In several variables, there are many different directions in which we can move.

Let $d \in \mathbb{R}^n$ be a direction. The *directional derivative* of f at x in the direction d is defined by

$$D_d f(x) = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t},$$

provided the limit exists.

If f is differentiable, then the directional derivative can be computed using the gradient:

$$D_d f(x) = \nabla f(x)^T d.$$

where $\nabla f(x)^T d$ is the dot product between the gradient and the direction vector. This tells us how we move from x in the direction d.

If d is a unit vector, then we have

- $\nabla f(x)^T d > 0$ means that f initially increases in the direction d.
- $\nabla f(x)^T d < 0$ means that f initially decreases in the direction d.
- $\nabla f(x)^T d = 0$ means that there is no first-order change in the direction d.

Critical Points

The idea of a critical point also extends naturally to several variables. A point x^* is called a *critical point* of f if

$$\nabla f(x^*) = 0$$

or if one or more of the partial derivatives do not exist at x^*.

Similar to the single-variable case, critical points are important candidates for local maxima and minima.

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at x^*, and suppose that f has a local minimum or local maximum at x^*. Then

$$\nabla f(x^*) = 0.$$

This is the multivariable version of the *first-order necessary condition* for optimality.

Similar to the one-variable case, this condition is necessary but not sufficient. The condition

$$\nabla f(x^*) = 0$$

does not by itself tell us whether x^* is a local minimum, local maximum, or neither

A critical point that is neither a local minimum nor a local maximum may be a *saddle point*

The Hessian Matrix

In single-variable calculus, we use the second derivative $f''(x)$ to study the curvature of a function. For functions of several variables, the corresponding object is the *Hessian matrix*.

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ has second-order partial derivatives. The Hessian matrix of f is

$$\nabla^2 f(x) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

For a function with two variables, this becomes

$$\nabla^2 f(x, y) = \begin{bmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{bmatrix}.$$

The Hessian plays the same role in multivariable calculus that the second derivative plays in single-variable calculus. It gives us information about the local curvature of the function.

Suppose x^* is a critical point satisfying

$$\nabla f(x^*) = 0.$$

The Hessian can often be used to determine what type of critical point x^* is

For now, it is important to know how to compute the hessian of a function. Later in the course, when learning unconstrained optimization we will discuss exactly what is important about the Hessian and how we use it.

Note: Here, we introduce a concept that you may or may not have seen in your linear algebra courses, but it is extremely important for optimization.

Analogous to the 1D case, in the multivariable case, the Hessian is used to deter,ine if a critical point is a local minimum, local maximum, or neither.

Let H be a symmetric matrix. We say that:

- H is *positive definite* if
$$x^T H x > 0$$
for every nonzero vector x
- H is *negative definite* if
$$x^T H x < 0$$
for every nonzero vector x
- H is *indefinite* if $x^T H x$ can take both positive and negative values for different choices of nonzero x
- Similarly, we say that a matrix is *positive semidefinite* if $x^T H x \geq 0$ and *negative semidefinite* if $x^T H x \leq 0$ for every nonzero vector x

  The difference between definiteness and semidefiniteness is that in the semidefinite case, it is possible to have
$$x^T H x = 0$$
  for some nonzero vector x

These ideas give us the multivariable version of the second derivative test

Suppose x^* is a critical point of f, so that
$$\nabla f(x^*) = 0.$$
Then:
- If $\nabla^2 f(x^*)$ is positive definite, then x^* is a strict local minimum.
- If $\nabla^2 f(x^*)$ is negative definite, then x^* is a strict local maximum.
- If $\nabla^2 f(x^*)$ is indefinite, then x^* is a saddle point.

This is directly analogous to the one-dimensional second derivative test. In one dimension, we look at the sign of
$$f''(x^*),$$
while in several dimensions we examine the definiteness of the Hessian
$$\nabla^2 f(x^*).$$

When the Hessian is *positive semidefinite* or *negative semidefinite*, the second-order test may be inconclusive, similar to the single variable case where
$$f''(x^*) = 0$$

We will discuss these ideas in more detail later in the course. For now, you should be comfortable computing the Hessian and understand what the definiteness says about the behavior of a function near a critical point.

Exercise.

Compute the gradient and hessian of the following functions, then evaluate their gradients at the given points respectively.

- $f(x, y) = x^2 + 3xy + y^2$, evaluate at $(5, 1)$
- $f(x, y) = x^4 - x^3y - xy^2 - y^3$, evaluate at $(3, 1)$
- $f(x, y, z) = x^2y + y^2z + z^2x$, evaluate at $(1, 2, 4)$

It is important to be comfortable with the computations above.

3 Lagrange Multipliers

So far we have seen how we find the minimum and maximum of functions when the function is free to take any value in the domain. However, several problems require us to maximize or minimize a function subject to some constraints.

For example, we may want to maximize or minimize

$$f(x, y)$$

subject to the condition

$$g(x, y) = c$$

The method of *Lagrange multipliers* gives us a way to find possible maxima and minima of a function subject to a constraint of the form $g(x, y) = c$

At a possible maximum or minimum, we look for points where

$$\nabla f(x, y) = \lambda \nabla g(x, y),$$

while also satisfying the original constraint.

> To maximize or minimize $f(x, y)$ subject to
>
> $$g(x, y) = c,$$
>
> we solve the system
>
> $$\nabla f(x, y) = \lambda \nabla g(x, y)$$
>
> together with
>
> $$g(x, y) = c.$$

Here, λ is called a *Lagrange multiplier*

Geometrically, the equation

$$\nabla f = \lambda \nabla g$$

means that the gradients of f and g are parallel at the point we are considering

This idea of Lagrange multipliers is extremely useful in Optimization and will appear frequently when we discuss nonlinear optimization later in the semester.

Example

Find the maximum and minimum values of

$$f(x, y) = xy$$

subject to

$$x^2 + y^2 = 1.$$

We first compute the gradients:

$$\nabla f(x, y) = \begin{bmatrix} y \\ x \end{bmatrix}$$

and

$$\nabla g(x, y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix},$$

where

$$g(x, y) = x^2 + y^2.$$

The Lagrange multiplier equations are therefore

$$y = 2\lambda x,$$

$$x = 2\lambda y,$$

together with the constraint

$$x^2 + y^2 = 1.$$

From the first two equations, we obtain

$$x = \pm y.$$

If $x = y$, then

$$2x^2 = 1,$$

so

$$x = y = \pm\frac{1}{\sqrt{2}}.$$

At these points,

$$f(x, y) = \frac{1}{2}.$$

If $x = -y$, then

$$x = \pm\frac{1}{\sqrt{2}}, \qquad y = \mp\frac{1}{\sqrt{2}},$$

and

$$f(x, y) = -\frac{1}{2}.$$

Therefore, the maximum value is

$$\frac{1}{2}$$

and the minimum value is

$$-\frac{1}{2}.$$

Exercises

Use the method of Lagrange multipliers to solve the following problems.

1. Find the maximum and minimum values of

$$f(x,y) = x + y$$

subject to

$$x^2 + y^2 = 1.$$

2. Find the maximum and minimum values of

$$f(x,y) = x^2 y$$

subject to

$$x^2 + y^2 = 1.$$

3. Find the maximum and minimum values of

$$f(x,y,z) = x + y + z$$

subject to

$$x^2 + y^2 + z^2 = 1.$$

4 Taylor Expansions

Taylor expansions allow us to approximate a function near a given point using a polynomial. This is very useful in optimization because simpler linear and quadratic approximations can help us understand how a function behaves locally. In a more advanced Optimization course, Taylor expansions are used extensively to study and prove convergence of numerical algorithms. In the multivariable setting, the first-order Taylor approximation uses the gradient and the second-order approximation uses the Hessian.

Single-Variable Taylor Expansion

Suppose that $f : \mathbb{R} \to \mathbb{R}$ is sufficiently differentiable near a point x_0. The n-th order Taylor expansion of f about x_0 is

$$f(x) \approx \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

Equivalently,

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n.$$

For optimization, we are particularly interested in the first-order and second-order approximations.

The first-order Taylor approximation is

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0),$$

while the second-order Taylor approximation is

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2.$$

The first-order approximation uses the slope of the function at x_0, while the second-order approximation also uses information about its curvature.

Multivariable Taylor Expansion

Taylor expansions also extend to functions of several variables. Suppose that

$$f : \mathbb{R}^n \to \mathbb{R}$$

is sufficiently differentiable near a point x_0.

The first-order Taylor approximation of f about x_0 is

$$f(x) \approx f(x_0) + \nabla f(x_0)^T(x - x_0).$$

The second-order Taylor approximation is

$$f(x) \approx f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0).$$

Notice that the first-order approximation uses the gradient, while the second-order approximation also uses the Hessian matrix.

The first-order Taylor approximation gives us a local *linear* approximation of the function, while the second-order Taylor approximation gives us a local *quadratic* approximation.